Adapting Security Warnings to Counter Online Disinformation

Ben Kaiser, Jerry Wei, Eli Lucherini, Kevin Lee, J. Nathan Matias, Jonathan Mayer

Princeton University

USENIX Security 2021
Outline

• Platform disinformation warnings: examples and evaluations
• Browser security warnings: a success story
• Our research: designing disinformation warnings that work
• Conclusions and recommendations
Why use warnings?

- **Add context**, instead of restricting speech
- **Induce resistance** to misbeliefs\(^1\) and increase susceptibility to corrections\(^2\)

\(^1\) Greene 1982 \(^2\) Ecker 2010
Research on Modern Warnings

Methods
- In-laboratory survey experiments
- Self-reported assessments of:
  - Perceived accuracy
  - Likelihood of sharing
- Contextual warnings only
  - Primarily “disputed” warnings

Findings
- Warnings modestly decrease perceived accuracy\([1,2,3]\)
- Prior exposure is more important than warnings\([3]\)
- 3 separate studies found that warnings had insignificant effects on accuracy judgments\([4,5,6]\)

1 Pennycook 2017  2 Clayton 2019  3 Pennycook 2018  4 Gao 2018  5 Ross 2018  6 Seo 2019
Browser Security Warnings

Goals

• Protect against phishing, MITM, malware, and other threats
• Retain user choice, which is important because of false positives

Research

• Clickthrough rate (CTR) is the key metric
• Early studies found high CTR (~70%)\textsuperscript{[1,2]}
• Methods evolved from surveys to supervised tasks to field studies
• Modern warnings achieve 10-25% CTR\textsuperscript{[3,4]}

Relevant findings

• Warnings must be noticeable, credible, and motivating
• Experimental tasks must be realistic
• Interstitials >> contextuas\textsuperscript{[1,5]}

Google’s interstitial warning for flagged sites [4].

\textsuperscript{1}Wu 2006  \textsuperscript{2}Schecter 2007  \textsuperscript{3}Akhawe 2013  \textsuperscript{4}Reeder 2018  \textsuperscript{5}Egelman 2008
Research Goals

Empirically evaluate **interstitial** and **contextual** disinformation warnings

- Will users **notice** and **understand** the warnings?
- Will users **change their behavior** after seeing the warnings?
- What **messaging strategies** are most effective at changing user behavior?
Qualitative Laboratory Study ($n = 40$)

**Methods:** think-aloud role-playing tasks & interviews
- 4 search tasks using Google Search & Chrome
- 2 control rounds & 2 treatment rounds, with either **contextual** or **interstitial** warnings
- **Primary** and **alternative** sources specified for each task

**Data**
- Researchers’ notes
- **Clickthrough rate** (CTR)
- A new metric: **alternative visit rate** (AVR)
- Follow-up interviews

The **contextual** warning (top) is adapted from the Google Search inline warning. The **interstitial** warning (bottom) is adapted from the Google Chrome SafeBrowsing warning.
Laboratory Results

Behaviors

Notice & Comprehension

Contextual ($n = 20$)
- 4 subjects did not notice the warning
- 9 more saw the icon but not the text

Interstitial ($n = 20$)
- 8 subjects did not realize the warning was about disinformation
- 7 of 8 still chose to go back

Takeaways

- The interstitial was noticeable, comprehensible, effective
- $\sim \frac{1}{2}$ of AVs were subjects who comprehended the warning
- 3 mechanisms of effect emerged:
  - Informativeness
  - Fear of harm
  - Friction
Quantitative Crowdworker Study

Goals

• Validate effect of interstitial warnings
• Identify informative & threatening designs and compare effect sizes
• Examine moderating effect of partisanship

Methods: search tasks & surveys (n = 238)

• 4 tasks using simulated search tool
• 8 warning designs; 4 for each theory of effect
• Treatments adaptively assigned
• Surveys after warning encounters
• Bonus payments for correct answers
• Track clicks to measure CTR & AVR

We adapted warnings from Google Chrome. For harm (top), we used the SafeBrowsing warning. For informativeness (bottom), we used the SSL warning.
Key Findings

- Participants were **significantly more likely** to visit alternative sources after seeing an interstitial warning
  - $z = 22.44$
  - $p < 0.001$
- Participants **reliably understood** our informative warnings
- Informativeness & harm scores had **no significant correlation** with AVR

<table>
<thead>
<tr>
<th></th>
<th>Liberal</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>AVR</td>
<td>CTR</td>
<td>i</td>
<td>h</td>
<td>#</td>
<td>AVR</td>
<td>CTR</td>
<td>i</td>
<td>h</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>318</td>
<td>20%</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>158</td>
<td>16%</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>318</td>
<td>87%</td>
<td>16%</td>
<td>–</td>
<td>–</td>
<td>158</td>
<td>85%</td>
<td>17%</td>
<td>–</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Selected treatments**

<p>| | | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>h1</td>
<td>120</td>
<td>85%</td>
<td>18%</td>
<td>–1.94</td>
<td>1.18</td>
<td>46</td>
<td>83%</td>
<td>17%</td>
<td>–1.91</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>h3</td>
<td>73</td>
<td>84%</td>
<td>18%</td>
<td>–</td>
<td>–</td>
<td>27</td>
<td>81%</td>
<td>22%</td>
<td>–</td>
<td>1.15</td>
<td>±0.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>i3</td>
<td>39</td>
<td>87%</td>
<td>13%</td>
<td>1.41</td>
<td>±0.43</td>
<td>–</td>
<td>10</td>
<td>90%</td>
<td>10%</td>
<td>–</td>
<td>–</td>
<td></td>
<td></td>
</tr>
<tr>
<td>i4</td>
<td>17</td>
<td>82%</td>
<td>12%</td>
<td>–</td>
<td>–1.76</td>
<td>±0.69</td>
<td>25</td>
<td>76%</td>
<td>24%</td>
<td>0.88</td>
<td>±0.69</td>
<td>–0.2</td>
<td>±0.62</td>
</tr>
</tbody>
</table>

Informativeness ( $\bar{i}$ ) and harm ( $\bar{h}$ ) scores are on an interval scale [-2,2]
Conclusions & Future Work

Conclusions
• Contextual warnings are easy for users to overlook
• Interstitial warnings can effectively communicate to users and change behavior
• Behavioral effects may not result from informed decision making

Future Work
• More behavioral research on disinformation warning effects
• Large-scale field studies
• Redoubled efforts, transparency, and cooperation by platforms
Acknowledgements

Collaborators: Jerry Wei, Eli Lucherini, Kevin Lee, J. Nathan Matias, and Jonathan Mayer

USENIX organizers, reviewers and our shepherd

Contact: bkaiser@princeton.edu